DOCUMENT RESUME

ED 314 441 TM 014 235

AUTHOR Schmitt, Dorren Rafael
TITLE The Validation of a Software Evaluation Instrument.
PUB DATE Nov 89
NOTE 28p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Little Rock, AR, November 8-10, 1989). Based on research for a dissertation, University of New Orleans.
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150) -- Dissertations/Theses - Doctoral Dissertations (041)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *College Students; *Computer Software; Construct Validity; Content Validity; *Courseware; Graduate Students; Higher Education; Predictive Validity; Secondary Education; *Secondary School Mathematics; Test Reliability; *Test Validity
IDENTIFIERS *Instrument for Software Evaluation for Educators; *Software Evaluation; Validation Verification and Testing Techniques

ABSTRACT

This study, conducted at six southern universities, analyzed the validity and reliability of a researcher developed instrument designed to evaluate educational software in secondary mathematics. The instrument called the Instrument for Software Evaluation for Educators uses measurement scales, presents a summary section of the evaluation, and provides recommendations for purchasing decisions. The study sample included 304 education students enrolled in graduate and undergraduate mathematics methods and computer education courses. The software was selected from commercial and public domain clearinghouses. A total of 52 packages were used on Apple and IBM personal computers. Content areas of the courseware included algebra, geometry, trigonometry, calculus, general mathematics, and statistics. Content, construct, and predictive validity as well as internal consistency and test-retest reliability of the evaluative instrument were investigated. Results indicate that the instrument has appropriate content validity, four stable constructs, and retest reliability. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED314411

The Validation of a Software Evaluation Instrument

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Dorren Rafael Schmitt

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Dorren Rafael Schmitt

University of New Orleans 70148

Based on Dissertation Research, May 1989

Major Professor: Charles Gifford, Ed.D.

University of New Orleans

Paper presented at the annual meeting of the Mid-South Educational Research Association, Little Rock, AR, November 1989.

TM014235

BEST COPY AVAILABLE

The Validation of a Software Evaluation Instrument

Educational software has become an integral part of educational plans over the past 15 years. A major increase in software prcduction and use began in the 1980s. During these years of software production, a proliferation of software evaluation instruments were developed. These instrument have many problems with them, from open-ended questions to questions assessing multiple aspects.

This study analyzed an instrument developed by this author. Several validity and reliability test were conducted to provide empirical evidence for the instrument. The study was conducted in six mid-sixed souther universities. The sample included 304 students enrolled in graduate and undergraduate mathematics methods and computer education courses from the colleges of education.

Content, construct and predictive validity, and internal consistency and test-retest reliability were investigated. Several conclusions were reached. Some of these were: (1) the instrument has appropriate content validity, (2) the instrument tests four stable constructs, and (3) the instrument has retest reliability.

The Validation of a Software Evaluation Instrument

The first computers were large mainframes built with vacuum tube technology in the 1940s only economically feasible for large corporations. Since the advent of the microcomputer in 1975, computer usage in elementary and secondary schools have escalated rapidly (Braun, 1981; Mathews & Winsauer, 1984; Wold & Hunter, 1984). In 1980, there were approximately 52,000 computers, for instructional purposes, in approximately 48% of the public school districts (Becker, 1986).

Since 1975, the availability of computers and the explosion of software for educational purposes has led to the development of numerous software evaluations. Dickey and Kherlopian (1987) reported that 75% of mathematics and science teachers in South Carolina have access to computers and 45% use computers for instructional purposes. Anderson (1987) confirmed the growth in the use of computers for instruction with a study that found the average high school used available computers approximately 2.5 hours per day.

With the increase in computer usage came evaluations of the software used with the computers. University and public school evaluations of educational software first emerged in the 1960s (Truette & Gillespie, 1984). In the late 1970s, around the same time educational software development increased at an exponential pace, large numbers of evaluation instruments were developed. The extensive marketing of microcomputer software spurred an increase of evaluations.

Formal evaluation of educational software is an important step in helping educators to decide, prior to purchase, about the usefulness and quality of the materials. Restricted school budgets dictate that educators be extremely careful about the software they select. Furthermore, instructional time is limited and ineffective software has the potential to just waste instructional time.

For the most part, the development of educational software evaluation instruments appears to be based on what the developer believes should be included in an effective program (Truette & Gillespie, 1984). The procedures involved in developing evaluation instruments appear to be: (1) assembling applicable concerns, (2) reviewing educational evaluation instruments designed for instructional materials other than software, (3) adding criteria specific to the new technology of computers, and (4) organizing these materials in some systematic fashion.

Initially, heuristics are necessary to establish the structure and the theory of a field (Caldwell, 1983). This is true for software evaluations as well. That is, the authors establish criteria based on their judgments of what "quality" software should be and what looks "good" on a computer screen. Caldwell (1983) argues "most of the criteria included on evaluation instruments are based on what the developer THINKS contributes to an effective program" (p. 5).

The problem with regard to the state of educational software evaluation is that hundreds of heuristically-based instruments

have been created and the field has no direction. Each instrument approaches software evaluation from a slightly different perspective. Too many heuristic efforts have **NOT** been followed by empirical research to determine which instrument has the greater validity and reliability. Mathews and Winsauer (1984) stated that it "would be ideal to set forth some standard by which software might be evaluated by any user" (p. 8). Van Muylwijk (1987) continued the idea of Mathews and Winsauer stating that it "seems to this point with all the varying instruments there is no 'ideal' evaluation instrument" (p. 453). Consequently, until the field moves away from many divergent, heuristically-based instruments to a single standard, there will be no "ideal" evaluation instrument achieved and consequently, educators will have little help in evaluating educational software before classroom adoption.

Validity and Reliability of Existing Instruments

Few efforts have been made to investigate either the validity or reliability of current software evaluation instruments. Up to the current effort, criterion validity and inter-rater reliability have been the only validity and reliability investigated in software evaluation instrument research. Kanselaar (1987) performed criterion validity testing between judges of software and inter-rater reliability. A multiple correlation was determined for each subscale and a Pearson correlation using 30 teachers and 23 software packages.

The multiple correlations achieved between judges on the two item categories were .66 and .61, and the inter-rater reliability was .52, statistically significant at the .05 level (Kanselaar, 1987).

Inter-rater reliability was also addressed by Cohen (1982) and Owston and Dudley-Marling (1986). Cohen (1982) attempted to establish inter-rater reliability with six individuals and software packages. Using a Pearson correlation, he found statistically significant results ($p<.05$). Owston and Dudley-Marling (1986) used the panel evaluation method with each panel containing two or three members. A Pearson correlation was performed with five panels to determine inter-rater (panel) reliability resulting in statistical significance ($p<.05$). Research in both the validity and reliability of software evaluation instruments is deficient. These previous studies have begun to address the problems, but more needs to be done.

A variety of criteria has been suggested by authors of instruments in the past decade. Bitter and Wighton (1987) and Mathews and Winsauer (1984) have surveyed instruments and discussed relevant common criteria of evaluation instruments. Other researchers (Bialo & Erickson, 1985; Marshall & Cannings, 1984) have performed the Delphi technique to ascertain appropriate criteria to be included in a software evaluation instrument.

Common criteria have been found in most software evaluation instruments. Bitter and Wighton (1987) surveyed the 28 members

of the Educational Software Evaluation Consortium in order to identify the 22 most common criteria. The criteria found by Bitter and Wighton (1987) included: correctness of presentation, content presentation, use of technology, integration into the curriculum, ease of use, curriculum congruence, interaction, content sequence levels, readability, user control of program, feedback, objectives, motivation, branching, negative feedback/help, content modification, content bias, teacher documentation, user support materials, color/sound/graphics, screen display, and management system. Mathews and Winsauer (1984) found 14 criteria appearing in all evaluations surveyed. Most of this criteria overlaps Bitter and Wighton (1987).

Marshall and Cannings (1984) found four categories of investigation for software evaluation instruments. Using the Delphi technique the categories of pedagogical content, instructional presentation, documentation, and technical adequacy were found to be the logical categories of items. Similar categories were derived by Bialo and Erickson (1985).

The primary purpose of this study was to develop an instrument for evaluating educational software in secondary mathematics that addresses several of the deficiencies of existing instruments (i.e., the lack of systematic software evaluation instruments, the lack of guidelines for decision making, and item wording). This new instrument, which was called Instrument for Software Evaluation for Educators (ISEE), uses measurement scales, presents a summary section of the evaluation,

and provides recommendations for purchasing decisions. It was expected that a formally developed instrument will advance the field of educational software evaluation. The ISEE was an attempt to provide decision-makers, such as teachers and administrators, with accurate information which will contribute to decisions about adopting software for educational purposes.

The second purpose of this study was to provide empirical evidence of validity and reliability for the ISEE. Initially, the ISEE is designed to evaluate secondary mathematics software packages. Thus, this study was limited to investigating the validity and reliability of ISEE relative to secondary mathematics software. The validity and reliability tests to be conducted include: content validity, construct validity, concurrent validity, internal consistency reliability, and test-retest reliability.

Methodology

Subjects

This study was conducted in six mid-sized southern universities. The sample included 304 students enrolled in graduate or undergraduate mathematics methods (48%) and computer education classes (52%). The majority of the students (70%) were graduate students.

The software was selected from commercial publisher and public domain clearinghouses. A total of 52 educational software packages were used for the Apple and IBM personal computers. The

content areas included were Algebra I and II, Geometry, Trigonometry, Calculus, General Math, and Statistics.

Procedures

Content validity was assessed through a literature review and two, five member expert panels. Each panel member evaluated the instrument independently.

The remaining questions were assessed using the 304 subjects with the 52 software packages. The researcher performed the study during one hour of a class period. The students reviewed and evaluated a software package during this hour. The computers were located at the university the students were attending, but the software was provided by the researcher.

The students were randomly seated at a computer with a software package and a copy of the evaluation instrument (ISEE). The students were instructed to read the introduction of the ISEE and use the software for approximately 45 minutes and fill out the ISEE during the last 15 minutes. Four weeks later, the researcher returned to a subsample of the classes (N=118 students), from professors allowing a second hour. These students in the retest evaluated the same package as previously evaluated.

Results

1. Does the instrument have content validity?

Initially, a content analysis was performed using 150

evaluation instruments to find the most common elements. The purpose was to take the most common and best items and refine them into an evaluation instrument to eliminate instrument design flaws found in previously designed evaluation instruments. Items incorporated also have previous empirical research indicating that these features of software were important to investigate.

The content analysis used was based on suggested methods for including items (Collins, 1974; Impellitteri & Kapes, 1971). The methods suggested included: (a) surveying existing instruments to find common items with supporting research, and (b) confirming item inclusion using an expert panel for content validity.

The program COMCO (Aiken, 1980a) was employed to produce a content validity coefficient between 0 and 1, with 1 indicating that all items had complete content validity for two expert panels(Aiken, 1980b). All members of both expert panels rated each item of the ISEE.

The first expert panel (panel 1) was composed of two faculty members from a local university, two Learning Resource Center (LRC) staff members, and one external to the university educational media consultant. The second expert panel (panel 2) was composed of two out-of-state university professors, a local educational media consultant, a media director, and a public school, junior high school teacher.

Both panels rated the ISEE high in content validity, with an average between the two panels of .898. The validity score, a correlation, for panel 1 was .892 and for panel 2 was .905. Each

panel member rated the ISEE at least .847 for content validity. Content validity coefficients were statistically significant at the .05 level (Aiken, 1980b).

Although the comprehensiveness and relevance of the criteria established in the ISEE could be questioned by other researchers, the evidence exists that the items included in the ISEE are valid for evaluating educational software. The literature review as well as related research investigating computer-assisted instruction (CAI) supported the inclusion of the items in the ISEE. The two expert panels empirically confirmed the literature review results.

The fact that ISEE has content validity implies that the ISEE evaluates what it purports to evaluate - educational software. To this point, this is the first study involving an educational software evaluation instrument to use both a literature review and empirical evidence to confirm content validity. Thus, an important step toward more empirical evidence of validity has been made in the development of evaluation instruments.

2. Does the instrument have construct validity?

The principle component factor analysis using a varimax rotation was performed on the 53 scale items from 304 participants. The factor analysis was utilized to determine the factor structure of the ISEE for construct validity.

The point of inflection on Cattell's scree plot (Cattell,

1966) and eigenvalues of at least 1.0 (Kaiser, 1958; Rummel, 1970) were used to determine the number of factors to extract. Based on the scree plot and the factor pattern coefficients, the best factor model was a four factor model. No factor extracted had an eigenvalue below 2.0 and all items had factor pattern coefficients, after rotation, of at least .40 (Stevens, 1986).

Because all items on the ISEE did not load highly on only one factor, the factor structure matrix was not in simple form. The eigenvalues variance, variance, and factor pattern coefficients for each of the factors are presented in Table 1.

INSERT TABLE 1 ABOUT HERE

Construct names were given to each factor based on qualitative analysis of the items factor pattern coefficients loading highly on a given factor and on comparisons of the factor names from the pilot study. The names for these factors include: Factor I - Performance Assessment and Software Impression (PAI), Factor II - Student/Software Individualized Environment (SIE), Factor III - Visual Displays (VD), and Factor IV - Motivating Learning and Creativity (MLC).

Additionally, Table 1 presents the communalities for each item and the factor pattern coefficients for each item. These communalities range from a high of .748 for item B17 to a low of .18184 for item D10. The communality is the portion of the variance that is accounted for by the common factor and is an estimate of the reliability of the item (Tatsuoka, 1988). Three items had low communalities: D1, D7, and D10. These three items

were not included in the factor structure matrix because of the low factor structure coefficients and the low communalities. Thus, with the factor analysis producing four factors, it seems that the ISEE actually has four subsections as opposed to three heuristically developed subsections.

Also the factors from the ISEE confirmed three of the broad categories previously found by Marshall and Cannings (1984) through the Delphi method. The categories include: (1) pedagogical content which matches with Factor I - PAI, (2) instructional presentation which matches with Factor II - SIE, and (3) technical adequacy which matches with Factor III - VD. The last category of Marshall and Cannings (1984), documentation, was not confirmed as an independent category, but was subsumed under Student/Software Individualized Environment. The fourth factor of the ISEE, MLC, was not previously found. The factor analysis provided limited support for previous research attempting to define global characteristics of educational software. These findings help to increase the understanding of what are the important characteristics of educational software.

3. What is the internal reliability of the subparts and the total instrument?

A Cronbach alpha was used to determine the internal consistency reliability for each of the subparts, the factors extracted, and the total instrument. The three subparts of the ISEE are: presentation, instruction, and content. The alpha for

presentation was .8220, for instruction was .9299, and for content was .8231. The alpha for the total instrument was .9453 (53 items). These alphas are acceptable for an instrument with four factors (Sax, 1980).

The alpha for each factor was also calculated with the items loading highly on that factor included in the factor alpha. The alphas for the factors were: .9466 (27 items) for factor I, .8221 (12 items) for factor II, .7488 (4 items) for factor III, and .7141 (7 items) for factor IV. Although two alphas were smaller than what might be desired, the low numbers of items and items loading on multiple factors attributed to the smaller alphas.

Based on the subpart, factor, and total instrument Cronbach alphas, the ISEE was internally consistent. The findings from the factor analysis affected which alphas were most important; they are those for the total instrument and the factors. But, the total score is what ultimate purchase decisions should be based upon and needs to be emphasized.

4. What is the test-retest reliability given four weeks lapse of time?

A Pearson product-moment correlation was used to determine test-retest reliability. Due to limitations of class time and availability of students, a subsample of the participants (N=118) was used to determine test-retest reliability. This subsample of students was from 3 universities in 5 classes (4 computer education and 2 mathematics methods). There were 45 males and 73

females.

The subscores and total scores of the first evaluation were correlated with the scores from the second evaluation, which was given four weeks later. All correlations were statistically significant. The Pearson correlation coefficients for the subtests and the total instrument from the first to second evaluation were .9551 ($p<.0001$) for presentation, .8779 ($p<.0001$) for instruction, .9710 ($p<.0001$) for content, and .9636 ($p<.0001$) for total.

Further analysis indicated that the differences between the first use total score and second use total score ranged from -.86 to .44 (negative implies that the second total score was higher than the first), with a mean and standard deviation of -.139 and .190, respectively. This indicates that there was little difference between the first and second use scores. Because of the differences being below one, the second total score could not have moved more than one decision level. The decision levels are the categories established in the ISEE introduction for evaluation/purchase determinations.

An analysis of change in the evaluation decision, based on the scores falling into one of four ranges for making an evaluation judgment, was investigated. The decision analysis indicated that 84 percent (99 of 118) of the total scores remained in the same range recommended in the introduction of the ISEE. The change in decision range was no more than ± 1 decision level based on the difference range. Because 90 percent of the

differences were negative, thus the decision level remained the same or increased by one.

Based on the correlation, the score differences, and the analysis of evaluation decisions, the ISEE seems to produce stable scores over a four week period. There is confidence that the ISEE is a stable evaluation of high school mathematics educational software.

5. What is the concurrent validity between the ISEE and other published evaluations?

An analysis with a subsample (N=17) of the software packages (all those packages in the study with published evaluations) was run to determine the concurrent validity. ISEE ratings were compared to evaluations from National Council of the Teachers of Mathematics (NCTM) in the Mathematics Teacher, from MicroSIFT, and from the Alberta Department of Education. The NCTM evaluations were scored by five local mathematics educators to provide a standard measurement across evaluations, since a standard scale was not used for NCTM evaluations.

The evaluations published by the Alberta Department of Education (Alberta) and NCTM do not specify the criteria used to evaluate the software packages. The evaluations are descriptive texts with broad conclusions and observations. For Alberta, a score between 1 and 5 is provided to give the reader a score for comparisons. NCTM also provides descriptions and broad conclusions for software packages. The NCTM evaluations do not

provide a score, therefore the five mathematics educators discussed above provided the evaluation ratings for this study.

The MicroSIFT published evaluations do provide some criteria that was used to evaluate the software. Additionally, descriptions of the software and broad conclusions are made. A evaluation score is provided by MicroSIFT for the purpose of comparing software packages and purchasing decisions. These evaluation may not be comparable to the ISEE, particularly NCTM and Alberta. Thus, at least NCTM and Alberta are the crudest of indices.

Table 2 presents the evaluation scores for each package evaluated. A difference, in evaluation scores between the ISEE and published evaluations, of less than ± 1 was found in any one evaluation in 15 of the 17 (88%) software packages. Additionally an average of the published evaluations was considered and in 12 of the 17 (71%) the difference was less than ± 1; thus the scores were similar. In three of the 17 cases all of the scores were greater than 1 score and in four cases the scores were similar to the ISEE but had at least one other evaluation score greater than a ± 1 difference. Difficulty in comparisons of the exists since the measures of quality are different (See notes in Table 2). Overall, there seems to be moderate concurrent validity since 88 percent of the evaluations were similar to published evaluations.

INSERT TABLE 2 ABOUT HERE

Additionally, Pearson correlations were run on the published evaluations, the means of those evaluations, and the ISEE. All

correlations, expect for MicroSIFT, were low with ISEE. The effect sizes (r^2) ranged from .09% for NCTM to 40.90% for MicroSIFT. NCTM and Alberta evaluations were highly correlated with each other. The higher correlation of the ISEE to MicroSIFT could be due to the fact it is more structured and has some common criteria with the ISEE.

The subparts of the ISEE were correlated with the NCTM, Alberta, and MicroSIFT evaluations. Both NCTM and Alberta were most highly correlated with Factor III .52 and .46, respectively. MicroSIFT was correlated with all of the subsection and the factors from .60 to .65. The only Factor III was not high correlated to MicroSIFT. NCTM and Alberta evaluations seem to only covers a small portion of the ISEE evaluation.

The differences between the ISEE and published evaluations in the evaluation scores may be due to several factors: (a) the ISEE provides a more precise measure of software quality, (b) many of the evaluators may not have been experienced in evaluating mathematics software, (c) the criteria used to evaluate the packages are not comparable, or (d) older, possibly different versions of the software were evaluated. Although "d" is unlikely, there is a slight possibility that a new version was developed for some of the software packages.

Although differences exist, the concurrent validity done was acceptable. Intervening variables aside, there seemed to be enough characteristics evaluated in common between the published evaluations and the ISEE to be sufficient so that any of the

software evaluations could be used. The question could be asked, which produces the most accurate ratings the ISEE or the published evaluations?

6. Is there a difference between educators using the ISEE having different educational evaluating experience? Is there a difference between educators using the ISEE with different software academic level?

To answer this research question, a 2 X 4 factorial analysis of variance (ANOVA) was employed. The dependent variable was the total score on the ISEE. The two independent variables were evaluation experience and academic level of the software. Evaluation experience had four levels (no experience - NE, experience evaluating other educational materials - EO, experience evaluating educational software - ES, and experience evaluating both other materials and software - EB). The academic level of the software had two levels (topics including general mathematics, statistics, and algebra - low level and topics including geometry, trigonometry,and calculus - high level). The ANOVA design was not balanced, but each cell contained at least 33 students; thus, there was more than sufficient power to find statistical significance.

No statistically significant differences were found in either the omnibus main effect hypotheses or in the interaction effect. Table 3 presents the ANOVA Summary Table for the analysis. The null hypotheses were not rejected, thus no post

hoc tests were performed.

INSERT TABLE 3 ABOUT HERE

The lack of statistical significance implies that there was no difference in the evaluation of software packages level. Additionally, the experience of the evaluator did not have a substantial effect on the evaluation of software using the ISEE. The expectation was that evaluation experience, especially with software, would have made a difference in the evaluation scores. But, it is possible that evaluation experience is not a reliable variable or one that accurately reflects the skills needed to evaluate software.

Conclusions

The ISEE has content validity. Based on the results from the literature review and the two expert panels, the ISEE has appropriate content validity for the purpose of evaluating educational software.

It seems that more than just the basic 22 common items found by Bitter and Wighton (1987) or the 14 items found by Mathews and Winsauer (1984) truly represent the full domain of items necessary for the evaluation of educational software. The ISEE contains 69 items (of which 53 are scale items) to evaluate educational software. All items were considered to have high content validity from two expert panels. Thus, the ISEE better represents the domain of items necessary for an accurate

evaluation of educational software than previous instruments have contained.

The ISEE was found to consists of four constructs. These four factors represent what seems to be a need to reorganize the ISEE. Instead of three subscales, there are four subscales. The factor analysis brought the structure of the ISEE from a heuristic perspective to one of an empirical perspective. The new structure requires modifications on the item groupings within the ISEE as well as a different weighting system.

Within each subpart and factor the ISEE was internally consistent. Weaknesses in Factors III and IV resulted in lower but acceptable internal consistency alphas. Overall, the items in the ISEE are homogeneous; they all focus on the main task of evaluating educational software and seem to do an acceptable job.

The ISEE has retest reliability. In 84 percent of the retest cases, the evaluation decision did not change. At least 90 percent of the retest case decisions remained the same or increased by one decision level.

The ISEE has concurrent validity. Scores on the ISEE were similar to evaluation scores from NCTM, MicroSIFT, and Alberta Department of Education. From a correlation of the three published evaluations, their means for each package, and the ISEE, the evaluation that most closely matched the ISEE was MicroSIFT. Of the three published evaluations, MicroSIFT evaluations are more extensive. The MicroSIFT evaluations are also well know and the most widely used of all published

evaluations. The correlations emphasize the differences in evaluations based on other instruments. But, the published evaluation based on one of the most comprehensive evaluation instruments most closely matched the ISEE evaluation.

The ISEE is a valid and reliable instrument that can be used by all mathematics educators regardless of evaluation experience. Although differences were expected, no statistically or educationally significant differences were found between educators experience levels. This indicates that no extensive evaluation training or experience is needed for an educator to use the ISEE. Additionally, the ISEE seems to exhibit stable scores across not only evaluation experience, but also academic level of software, and the interaction between evaluation experience and level of software. Thus, software bias in not a factor.

Overall, the ISEE does discriminate between poor software and good to excellent software. Through the extensiveness of the ISEE and the instrument design, ISEE can provide educators a valid and reliable instrument to evaluated educational software. This is the first software evaluation instrument to furnish extensive empirical data supporting claims of validity and reliability that many existing instruments imply having, but have avoided assessment. The results from this study alone provide the users of the ISEE with some knowledge of the weight that can be placed on the evaluation scores derived from the ISEE.

References

Aiken, L. R. (1980a). Comco [Computer Program]. Social Science Division: Pepperdine University Malibu: Malibu, CA 90265.

Aiken, L. R. (1980b). Content validity and reliability of single items or questionnaires. Educational And Psychological Measurement, 40, 955-959.

Anderson, R. E. (1987). The unresolved need for research on computers and education. Education and Computing, 3, 15-20.

Becker, H. J. (1986). Our national report card: Preliminary results from the New John Hopkins Survey. Classroom Computer Learning, 6(4), 30-33.

Bialo, E. R., & Erickson, L. B. (1985). Microcomputer courseware: Characteristics and design trends. AEDS Journal, 18, 227-236.

Bitter, G. G, & Wighton, D. (1987). The most important criteria used by the educational software evaluation consortium. The Computing Teacher, 14(6), 7-9.

Braun, L. (1981). Computer-aided learning and the microcomputer revolution. Programmed Learning and Educational Technology, 18, 223-228.

Caldwell, R. (1983). Evaluation of microcomputer software: How valid are the criteria and procedures? Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX. (ERIC Document Reproduction Service No. ED 230 171).

Cattell, R. B. (1966). The scree test for the number of factors. Multivariate Behavioral Research, 1, 245-276.

Cohen, V. L. (1982). Evaluating instructional software for the microcomputer. Paper presented at the annual meeting of the American Educational Research Association, New York. (ERIC Document Reproduction Service No. ED 216 704).

Collins, T. R. (1974). The importance of assessment instrument development. (ERIC Document Reproduction Service No. ED 107 729).

Dickey, E., & Kherlopian, R. (1987). A survey of teachers of mathematics, science, and computers on the use of computers in grades 5-9 classrooms. Educational Technology, 27(6), 10-14.

IEEE (1983). IEEE standard glossary of software engineering terminology. IEEE std 729.

Impellitteri, J. T., & Kapes, J. T. (1971). The measurement of occupational values. (ERIC Document Reproduction Service No. ED 059 357).

Kaiser, H. F. (1958). The varimax criterion for analytic rotations in factor analysis. Psychometrika, 23, 187-200.

Kanselaar, G. (1987). The description and evaluation of computer based instruction. In J. Moonen and T. Plomp (Eds.), Eurit 86 developments in educational software and courseware (pp. 445-449). Oxford: Pergamon Press.

Marshall, R. W., & Cannings, T. R. (1984). Evaluating educational software: A comprehensive model. Paper

presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Mathews, J., & Winsauer, J. (1984). Computer software for vocational education: Development and evaluation. State-of-the-art paper. (ERIC Document Reproduction Service No. ED 254 619).

Owston, R. D., & Dudley-Marling, C. (1986). A criterion-based approach to software evaluation. Paper presented at the annual meeting of the American Educational Research Association San Francisco, CA. (ERIC Document Reproduction Service No. ED 267 122).

Rummel, R. J. (1970). Applied factor analysis. Evanston, Il: Northwestern University Press.

Tatsuoka, M. M. (1988). Multivariate analysis techniques for educational and psychological research (2nd ed.). New York: Macmillan.

Truette, C., & Gillespie, L. (1984). Choosing educational software a buyer's guide. Littleton, CO: Libraries Unlimited.

Van Muylwijk, B. (1987). What to do with evaluation of educational software? In J. Moonen and T. Plomp (Eds.), Eurit 86 developments in educational software and courseware (pp. 451-455). Oxford: Pergamon Press.

Wold, A. L., & Hunter, C. B. (1984). The not quite twenty-year history of educational software. Media and Methods, 37, 24-25.

Table 1
Factor Structure Matrix after Varimax Rotation for the ISEE

Item	Communality	PAI	SIE	VD	MLC
Eigenvalues		16.113	3.225	3.040	2.364
Variance		30.4%	6.1%	5.7%	4.5%
ITMC13	.63564	.74226			
ITMC14	.63592	.71611			
ITMC1	.67754	.70959			
ITMC4	.57657	.69922			
ITMC6	.49597	.69565			
ITMD8	.54147	.67840			
ITMC7	.46915	.67780			
ITMC5	.61658	.65923			
ITMD9	.52731	.65016			
ITMB21	.49135	.62790			
ITMC3	.56529	.62060			
ITMD11	.51567	.61625			
ITMD12	.52186	.60299			
ITMC12	.58612	.60145			
ITMC8	.39457	.60074			
ITMB20	.44338	.58840			
ITMC17	.62020	.57304		.43350	
ITMD4	.37934	.57018			
ITMC11	.54974	.56104			.45457
ITMB16	.45798	.54657			
ITMB22	.53055	.53800			
ITMC2	.42687	.50380			
ITMD6	.44526	.50336			
ITMC10	.43994	.44281			
ITMC9	.32091	.42928			
ITMD2	.32155	.40946			
ITMB13	.36227	.40399			
ITMD10	.18184				
ITMD7	.23956				
ITMB4	.45988		.67095		
ITMB1	.52956		.62698		
ITMB9	.51970		.61635		
ITMB7	.50107		.61006		
ITMB8	.43813		.58465		
ITMD13	.52073		.56685		
ITMD3	.52455		.55791		
ITMB5	.35492		.45782		
ITMB3	.32515		.43811		
ITMB2	.29065		.41621		
ITMB15	.21687		.41540		
ITMB12	.21268		.40854		
ITMD1	.20291				
ITMB17	.74785			.84329	
ITMB18	.73305			.83645	
ITMB10	.72209			.81042	
ITMB14	.35576			.40612	
ITMB11	.40637				.60770
ITMB19	.42407				.59130
ITMC18	.52420				.55647
ITMC16	.68186	.55217			.55453
ITMB6	.31683				.54499
ITMC15	.41458	.40093			.42676
ITMD5	.34684				.41296

Table 2

Evaluation Scores for Each Software Package compared with Three Published Evaluations

Software Title	NCTM[a]	MicroSIFT[b]	Alberta[c]	ISEE
Algebra Drill & Practice I	3.6		3	2.92
Algebra Drill & Practice II		4	3	4.05
Basic Statistical Concepts	2.4			4.06
CactusPlot	3.8			3.78
Calculus Graph	4.0			4.39
Expert Algebra Tutor	4.0			3.98
Expression Writer	3.4			2.87
Geometric Supposer: Quad	4.2			3.42
Geometric Supposer: Tri	4.4		5	3.65
Graphing Linear Functions	4.0		2	4.01
Green Globs & Graph Equat.	4.2	4	4	3.77
Interpreting Graphs		4		4.03
Recycling Logic	3.2			3.84
Simultaneous Equations			4	4.17
Solving Fractional Equat.	4.5		4	4.60
Super Factory	5.0	4	5	3.34
Trap-A-Zoid		3		4.33

Note. [a]NCTM Scale: 1-Poor, 2-Fair, 3-Good, 4-Very Good, 5 - Excellent
[b]MicroSIFT Scale: 1-low to 5-high
[c]Alberta Scale: 1-Poor, 2-Fair, 3-Good, 4-Very Good, 5-Excellent

Table 3

ANOVA Summary Table for the Effect of Software level and Evaluation Experience on the Total ISEE score

Source	SS	df	MS	F	r^2
Main Effect					
Software Level (SL)	.518	1	.518	.340	.003
Evaluation Experience (EE)	2.587	3	.862	.209	.015
Interaction Effect					
SL X EE	4.450	3	1.483	.051	.025
Residual	167.590	296	.566		

Note. Homogeneity of Variance was calculated by the Bartlett-Box and was not statistically significant at the .05 level.